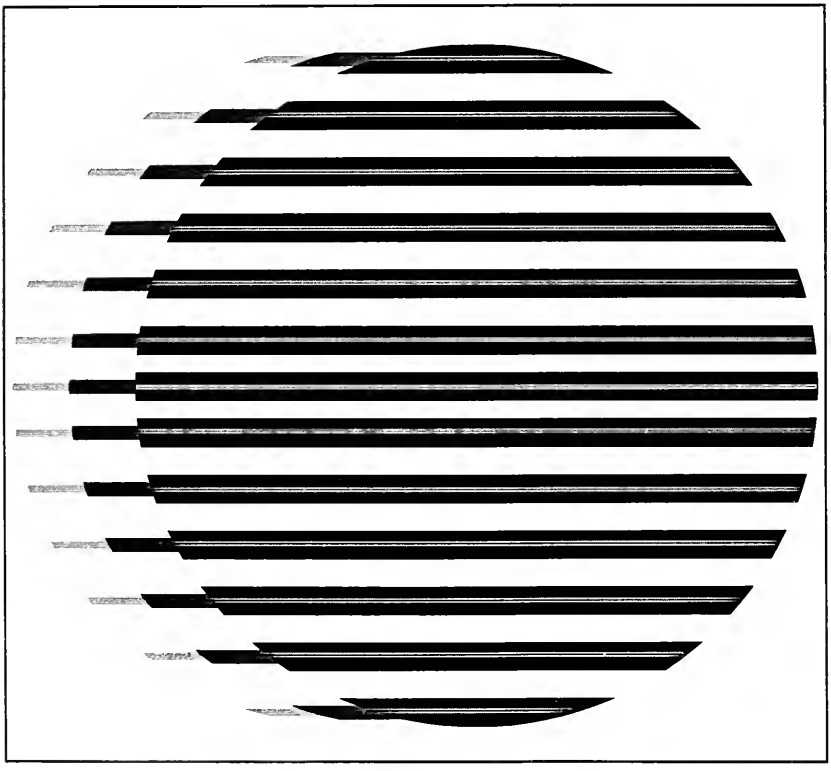
IASSIST

QUARTERLY

VOLUME 20 FALL 1996 NUMBER 3



Printed in the U.S.A.

IASSIST QUARTERLY

The **IASSIST QUARTERLY** represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The **QUARTERLY** reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of **IASSIST**.

Information for Authors

The **QUARTERLY** is published four times per year. Articles and other information should be typewritten and double-spaced. Each page of the manuscript should be numbered. The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". *Journal of the American Society for Information Science 30(2):77-82, March 1979*. If the contribution is an announcement of a conference, training session, or the like, the text should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event. Book notices and reviews should not exceed two double-spaced pages. Deadlines for submitting articles are six weeks before publication. Manuscripts should be sent in duplicate to the Editor: Laura Bartolo, Libraries & Media Services, Kent State University, Kent, Ohio 44242. (216) 672-3024 Email: LBARTOLO@KENTVM.KENT.EDU. Book reviews should be submitted in duplicate to the Book Review Editor: Daniel Tsang, Main Library, University of California P.O. Box 19557, Irvine, California 92713 USA. (714) 856-4978 E-Mail: DTSANG@ORION.CF.UCI.ED

Title: **Newsletter - International Association for Social Science Information Service and Technology**
ISSN - United States: 0739-1137

CONTENTS

Volume 20 **Number 3** **Fall 1996**

FEATURES

Disseminating Historical Census Data on the World Wide Web

by Steven Ruggles, Matthew Sobek, and Todd Gardner[1], University of Minnesota, Department of History

Introduction

This paper describes our project to electronically disseminate the Integrated Public Use Microdata Series (IPUMS). The IPUMS—the world's largest publicly available demographic database—is a coherent series of individual-level census data drawn from eleven census years between 1850 and 1990. Prior to the IPUMS, the U.S. census samples were a haphazard assortment of files created by different researchers at different times, each with its own unique record layout and coding schemes. By putting all the census samples in a compatible format, the IPUMS greatly simplifies the use of multiple census years. The database constitutes our most powerful resource for the study of long-term American social and economic change.

The development of the Internet and World Wide Web (WWW) promises to transform fundamentally the nature of electronic data dissemination. At the same time, the proliferation of fast personal computers and UNIX workstations is already revolutionizing data analysis. Our project capitalizes on both of these developments by making the largest and most powerful population database readily available for analysis on desktop machines.

The scholarly community has already shown great interest in the IPUMS. Nevertheless, the sheer size of the database poses problems for researchers. The Internet provides the most practical means of dissemination, but current methods of data distribution on the Internet are limited. Although enormous computational resources are becoming available to researchers, access to large-scale microdata is still cumbersome and expensive. The logistical difficulties are compounded by the spotty to nonexistent support for any but the most recent census samples.

We are currently in the process of implementing a project that addresses these distribution and usability issues. It is composed of three complementary elements, the second of which we have not yet undertaken:

1. Development of a data extraction system for use on the WWW. The preliminary version, described in detail below, is already in place. Users can fashion subsamples containing only those years, subpopulations, and variables that suit their research interests and computing power. In the future, we envision a Java-based extract system that is truly interactive, incorporating a variety of advanced features enabling researchers to customize their data extracts. They will also be able to construct new variables that capitalize on the hierarchical structure of the database.

2. Conversion of the IPUMS documentation into hypertext format to facilitate navigation. Users will be able to jump instantly to relevant sections of the documentation with the click of a mouse without negotiating 3000 pages of text. The documentation will be integrated with the extract system so that researchers can make informed choices in designing their subsamples. By using Adobe Acrobat format, the documentation will be downloadable onto virtually any computer platform while retaining its hypertext functions.

3. Ongoing support for users. For the first time, all of the existing census samples will be supported, in their integrated format.

The IPUMS has the capacity to become a cornerstone in the infrastructure for social science research. Our project is intended to remove the remaining obstacles preventing a wide range of researchers from taking advantage of the unique resource that the census samples represent. We hope our project can serve as a model for the distribution of microdata more generally.

Context

Since 1978, the federal government has invested approximately $19 million (in 1995 dollars) to create historical Public Use Microdata Samples (PUMS) of the decennial censuses for the period 1850 to 1950 (Graham 1980; Strong 1989; Ruggles and Menard 1994; Ruggles et al 1995; U.S. Bureau of the Census 1984a, 1984b). Beginning with the 1960 census, the Census Bureau has produced PUMS as a byproduct of each decennial enumeration (U.S. Bureau of the Census 1972, 1973, 1982, 1992). We now have a series of microdata for eleven census years (1850, 1880, 1900, 1910, 1920, 1940, 1950, 1960, 1970,

1980 and 1990), and a proposal to create two more samples (for 1860 and 1870) was recently funded by NICHD.

Taken together, these data files comprise our most powerful resource for the study of historical social and economic change. The range of potential topics that can be addressed with the national census files includes household composition, fertility, life-course transitions, ethnicity, immigration, internal migration, female labor force participation, the household economy, industrial and occupational structure, urbanization, nuptiality, and education. High-precision sample designs allow national and regional estimates that are virtually as reliable as published census data, but which have far greater subject area coverage. As microdata, rather than aggregate summary data, the samples provide information about individual persons and households transcribed directly from the original census manuscripts. The microdata contain far richer information than was ever published in the census volumes. They enable researchers to make tabulations tailored to their specific research questions and to overcome incompatibilities in the published census tabulations. In addition, they have allowed researchers to move beyond simple tabular analysis and apply increasingly sophisticated multivariate techniques. Although most of these census files have only been available for a few years, they have already led to an outpouring of new research (e.g., Jacobs 1989; Hirschman and Kraly 1990; Mare 1991; Jenson 1991; Kalmijn 1994; Ruggles 1994a, 1994b; Watkins 1994; Gjerde and McCants 1995).

The national census files have three key strengths: complete geographic coverage, large sample populations, and broad chronological scope. Complete geographic coverage is important not only because it allows scholars to generalize at the national level; national samples can also provide context for local studies. Moreover, by linking the census microdata to aggregate sources describing local characteristics, the PUMS allow multi-level analyses of the effects of local conditions on individual and family behavior.

The second strength of the national public use census files is their large size. The number of cases available for each census year ranges from the hundreds of thousands to the tens of millions. This allows the study of small and geographically dispersed population subgroups. For example, researchers at the University of Minnesota using the historical public use samples have examined topics such as the professionalization of nursing, American Indian fertility patterns, the living arrangements of elderly urban blacks, the demography of the prison population, the gender composition of clerical workers, and the living arrangements of parentless children. These research topics could not be pursued using a general social survey of the scale ordinarily undertaken by academic social scientists. Indeed, even the largest social survey carried out by the government—the Current Population Survey—is far too small for the detailed analysis of topics such as American Indian fertility or the professionalization of nursing (Olson 1991; Shoemaker 1991). The public use samples are the only general source of microdata available for any period with sufficient cases to study such small population subgroups.

The third, and most important, strength of the historical public use census files is their potential for the study of social and economic change over long periods of time. There is no other consistent source of quantitative information about the American population spanning more than a few decades. Despite frequent changes in subject content and modifications of enumeration procedures, the core of the census has remained remarkably stable over the past century and a half (Magnuson 1995). Since 1850, every census consists of a listing of individuals within households in a prescribed sequence and provides data on basic demographic characteristics such as age, sex, race, and birthplace. Table 1 indicates the broad range of subject areas covered in each census year.

Although the PUMS files are the most widely used data in American social science, few researchers have exploited the great potential of the national census files for the study of change over time. Instead, most investigators use single samples as isolated cross-sections (e.g., Haines 1989; Johnson and Lean 1985; Sanderson 1987; Sandefur and Sakamoto 1988; Sorenson 1989; Gordon and McLanahan 1991; Morgan et al 1993; Farley and Frey 1994; Krivo 1995; Sassler 1995). This is mainly because of incompatibilities among the original samples. The PUMS were created at different times by different investigators and, as a result, they have incompatible documentation and a wide variety of record layouts and coding schemes. Several previous census microdata projects—the samples of 1900, 1910, 1940 and 1950—ran out of money before they were finished. In most cases, the basic data were fine, but compromises were made in dissemination, user support, and documentation. Indeed, in the cases of the 1940 and 1950 public use microdata samples, critical sections of documentation essential for the proper use of the samples were omitted.

To resolve the incompatibilities among census samples and limitations of their documentation, NSF funded a project entitled"Integrated Public Use Microdata Series." The project incorporated all 23 existing national census samples into a single coherent database with an integrated set of documentation (Ruggles and Sobek 1995). The constituent samples are described in Table 2. All census years receive the same record layout and coding, without any loss of information from the original PUMS. Missing data in the early census years is imputed for the more important demographic variables. The

Table 1. Availability of Select Subject Areas Across Census Years

	185	1880	1900	1910	1920	1940	1950	1960	197	1980	1990
Household Record											
State	X	X	X	X	X	X	X	X	X	X	X
County	X	X	X	X	X	.	.	.	.	.	.
County group/public use microdata area	.	.	.	.	.	.	.	.	X	X	X
State economic area	X	X	X	X	X	X	X	.	.	.	.
Metropolitan status	X	X	X	X	X	X	X	X	X	X	X
Metropolitan area	X	X	X	X	X	X	X	.	X	X	X
City	X	X	X	X	X	X	X	.	.	X	X
Size of place	X	X	X	X	X	X	X	.	.	X	X
Urban/rural status	X	X	X	X	X	.	.	X	X	X	X
Farm	X	X	X	X	X	X	X	X	X	X	X
Ownership of dwelling	.	.	X	X	X	X	.	X	X	X	X
Mortgage status	.	.	X	X	X	.	.	.	.	X	X
Value of house or property	.	.	.	.	.	X	.	X	X	X	X
Monthly rent	.	.	.	.	.	X	.	X	X	X	X
Total family income	.	.	.	.	.	.	X	X	X	X	X
Person Record											
Relationship to household head	X	X	X	X	X	X	X	X	X	X	X
Age	X	X	X	X	X	X	X	X	X	X	X
Sex	X	X	X	X	X	X	X	X	X	X	X
Race	X	X	X	X	X	X	X	X	X	X	X
Marital status	.	X	X	X	X	X	X	X	X	X	X
Age at first marriage	.	.	.	.	.	X	.	X	X	X	.
Duration of marriage	.	.	X	X	.	.	X	.	.	.	.
Times married	.	.	.	X	.	X	X	X	X	X	.
Children ever born	.	.	X	X	.	X	X	X	X	X	X
Birthplace	X	X	X	X	X	X	X	X	X	X	X
Parents' birthplaces	.	X	X	X	X	X	X	X	X	.	.
Ancestry	.	.	.	.	.	.	.	.	.	X	X
Years in the United States	.	.	X	X	X	.	.	.	X	X	X
Mother tongue	.	.	.	X	X	X	.	X	X	.	.
Language spoken	.	.	.	X	.	.	.	.	.	X	X
School attendance	X	X	X	X	X	X	X	X	X	X	X
Educational attainment	.	.	.	.	.	X	X	X	X	X	X
Literacy	X	X	X	X	X	.	.	.	.	.	.
Employment status	.	.	.	X	.	X	X	X	X	X	X
Occupation	X	X	X	X	X	X	X	X	X	X	X
Industry	.	.	.	X	X	X	X	X	X	X	X
Class of worker	.	.	.	X	X	X	X	X	X	X	X
Weeks worked last year	.	.	.	.	.	X	X	X	X	X	X
Weeks unemployed	.	X	X	X	.	X	X	.	.	.	.
Total personal income	.	.	.	.	.	.	X	X	X	X	X
Wage and salary income	.	.	.	.	.	X	X	X	X	X	X
Migration status	.	.	.	.	.	X	X	X	X	X	X
Veteran status	.	.	.	X	.	X	X	X	X	X	X
Name	X	X	.	.	X	.	.	.	.	.	.

database also includes a series of fully compatible constructed variables. For example, the IPUMS provides "pointer" variables identifying the position within the household of every individual's own mother, father, and spouse. These and other variables, constructed identically for all years, provide the building blocks for researchers to design their own variables. Documentation includes comparability discussions for every variable as well as separate essays on the more complicated aspects of the data series.

We began distributing a preliminary version of the IPUMS data in September 1993 through an anonymous FTP site, and in April 1995 added a World Wide Web site. User response to our early data products has been overwhelming. When we set up our site in 1993 to distribute beta-test copies of a preliminary version of the IPUMS, our goal was simply to obtain feedback from experienced researchers on the design of the database. We expected to distribute only a few copies of the data, but news of its availability on the Internet spread quickly through the research community. Figure 1 gives the volume of IPUMS data downloaded by outside researchers since the preliminary release in late 1993.

Interest in the IPUMS database among social scientists clearly is high. In March 1995, the University of California at Riverside sponsored an All-Campus University of California Economic History Conference that drew fifty current and prospective users of the IPUMS to discuss methodological issues and present early research results from the database. The population centers at the Universities of Michigan, Wisconsin, Texas, Minnesota, SUNY-Buffalo, and Chicago have invited us to describe the data and its applications to interested researchers and data archive staff. In addition we have carried out extensive correspondence with early users of the IPUMS data. We designed the project in response to input from all these sources.

Despite the intense interest in the IPUMS, many researchers still have problems managing such large files. The number of scholars who have made effective use of the data so far represent only a small minority of those potentially interested. Although a large number of researchers have accessed the data, after downloading a file or two many have realized they lacked the resources to manipulate and analyze the data. The IPUMS contains approximately 69 million records, spans 140 years, and incorporates 524 separate variables. The sheer size of the IPUMS database (25 gigabytes) presents new challenges for data distribution (see the last column of Table 2). The Internet provides the most practical means of dissemination, but current methods of Internet data distribution are limited. Storage and distribution issues are the most consistent complaints conveyed by researchers using the PUMS and IPUMS. Most users need to decompress the data in order to use it, but computers with sufficient storage to decompress even the 1-in-100 samples are rare. The development of distributed computing environments has placed enormous computational resources in the hands of researchers, but access to large-scale microdata is still cumbersome and expensive. Even the documentation for the IPUMS database suffers from size problems: at 3000 pages, it is almost as unwieldy as the data itself.

To the extent that poor documentation, inadequate dissemination, and limited user support have curtailed use of the microdata samples, the resources invested in the PUMS samples have been wasted. The IPUMS project has the potential to correct these deficiencies, but only if access and use can be simplified. The significance of our project is not limited to the support and dissemination of the IPUMS, however. We hope to provide a model for the distribution of microdata generally. Accordingly, we will make our extract program and hypertext interface available to all interested researchers.

Extract System Design

To address the aforementioned distribution and usability issues, we have developed an extract system for use on the WWW that allows researchers to select only those subpopulations and variables needed for a particular analysis. In the future, we will integrate the documentation with the extract system and present it in hypertext format to facilitate navigation. In addition, we will offer user support of all the existing census samples in their IPUMS format. Each of these three aspects of the project is described in detail below.

1. Interactive extract system

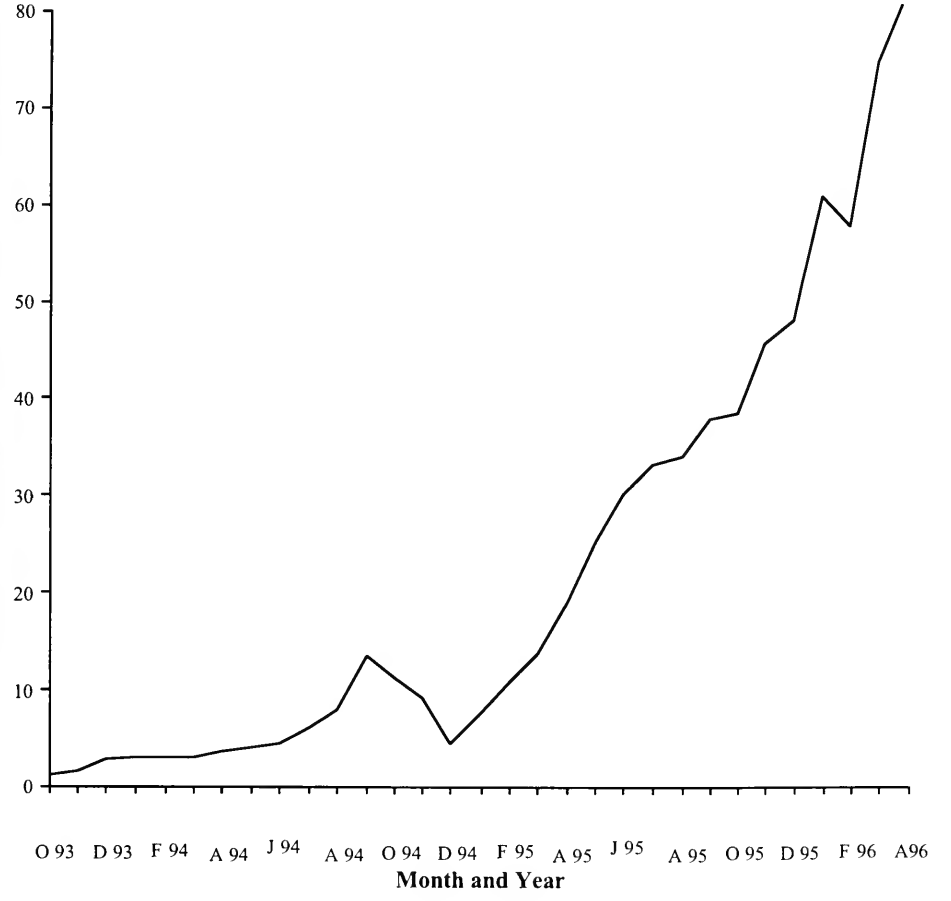
The great size of the census microdata files has always been a major obstacle to their use. Accordingly, we have developed an interactive extract system on the World Wide Web to provide easy access to the IPUMS from personal computers, workstations, and mainframe computers. The system allows researchers to fashion smaller extracts of the data specifically oriented to their own research needs and suited to their available computing power and storage capacity. In practice, researchers never require all variables and all cases from a census year. In the past, however, they have had no choice but to obtain the entire census samples to get the cases they wanted. Consequently, innumerable gigabytes of unused data are occupying tapes, hard drives, and other storage media across the country. With our extract system, researchers can design subsamples incorporating a subset of variables pertaining to the specific population(s) of interest to them.

Table 2. Descriptions and Sizes of the Public Use Microdata Samples Incorporated in the IPUMS

Sample Description	Year Released	Sample Density	Number of Records (thousands) Household	Person	Number of Variables	File Size
1850 PUMS -- Free population	1994	1 in 100	37	198	92	79 Mb
1860 PUMS -- General sample*	2001	1 in 100	66	354	94	141 Mb
1870 PUMS -- General sample*	2001	1 in 100	80	428	94	170 Mb
1880 PUMS -- General sample	1994	1 in 100	107	503	123	204 Mb
1900 PUMS -- General sample	1980	1 in 760	27	100	94	43 Mb
1910 PUMS -- General sample with Hispanic and Black oversamples*	1996	varies	113	480	125	198 Mb
1920 PUMS -- General Sample	1998	1 in 100	257	1037	122	433 Mb
1940 PUMS -- General sample	1984	1 in 100	391	1351	174	584 Mb
1950 PUMS -- General sample	1984	1 in 100	461	1922	170	798 Mb
1960 PUMS -- General sample	1971	1 in 100	579	1780	141	790 Mb
1970 PUMS -- 5% State sample	1972	1 in 100	744	2030	206	929 Mb
1970 PUMS -- 15% State sample	1972	1 in 100	744	2030	210	929 Mb
1970 PUMS -- 5% County group sample	1972	1 in 100	744	2030	203	929 Mb
1970 PUMS -- 15% County group sample	1972	1 in 100	744	2030	207	929 Mb
1970 PUMS -- 5% Neighborhood sample	1972	1 in 100	744	2030	260	1016 Mb
1970 PUMS -- 15% Neighborhood sample	1972	1 in 100	744	2030	264	1016 Mb
1980 PUMS -- "A" Sample	1983	1 in 20	4711	11337	276	5376 Mb
1980 PUMS -- "B" Sample	1983	1 in 100	942	2267	276	1075 Mb
1980 PUMS -- "C" Sample	1983	1 in 100	942	2267	266	1075 Mb
1990 PUMS -- 5% Sample	1992	1 in 20	5528	12500	252	6039 Mb
1990 PUMS -- 1% Sample	1992	1 in 100	1106	2500	252	1208 Mb
1990 PUMS -- 3% Elderly sample*	1993	1 in 33	*	*	*	*
1990 PUMS -- 1% Unweighted sample	1995	1 in 100	1106	2500	252	1208 Mb
					TOTAL	25,171 Mb

* = not yet in the IPUMS. At present, the IPUMS contains the 1910 PUMS general sample and a preliminary version of the 1920 PUMS.

Figure 1. Volume of IPUMS Data Downloaded: Gigabytes per month, three-month moving average



Our chief goal in designing an extract interface was to simplify access to historical census microdata, a task complicated by the sheer size and complexity of the database and the differing availability of variables across samples. The challenge was designing a system that makes these complications invisible to the user.

a. Preliminary version of the interface

We have already resolved some of the most difficult design issues for developing a new generation of data extraction software. The key component is the user interface, which was developed using the programming language Perl. This is still in preliminary form, but the design incorporates many of the features we envision. Before users initiate a data extract, they are prompted for their e-mail address, which provides us with a means of contacting them and constructing a unique file name for their extract output.

The extract procedure involves four steps—each on a separate Web page—with the contents of each page depending on selections made on the previous page. In the future, at any stage of the procedure, a query button will provide context-sensitive help explaining in detail all of the choices available to the user.

On the first page, partially shown in Figure 2, users define the general characteristics of their desired extract. They select the particular census sample or combination of samples they want (e.g., 1970 5% state sample, or the 1880 general sample) and the preferred file structure for their extract: hierarchical (household record followed by person records) or rectangular ("flat"—all household information attached to respective household members). Several sample densities are available, ranging from 1-in-20 samples available in recent census years to very small ("tiny") samples constructed in all years for purposes of testing and instruction. A feature allowing continuously variable sample densities will be added in the future. Finally, researchers may elect to include data quality flags, in which case the program will automatically append the flags corresponding to selected variables.

In the example shown in Figure 2, the user has elected to extract from the full ("regular") samples for 1850, 1880, 1950, and 1980 B. The extract will be produced in hierarchical format and will include data quality flags.

On the second page of the extract interface users select which variables they want to include in their extract. Only those variables available for the particular samples selected on the first page are displayed as options. If users have selected multiple census samples, all variables occurring in any of the specified samples are available. In all cases the form briefly describes each variable and indicates its availability among samples. Some variables have a second check box allowing users to select cases based on the value of the variable. In the future, we will add case selection boxes for many more variables. In addition, clicking on a variable name will call up all relevant documentation (see below). Users can also select entire groups of related variables by checking a single box at the end of each variable group.

Figure 3 partially displays the selections available to a user based on the choices entered in Figure 2. Only the samples for 1850, 1880, 1950, and 1980 are shown, along with the variables available in each of those years. In this case, the user has chosen a set of basic demographic variables (checked in the left-hand column). In the case of age, sex, race, and birthplace, the user has checked the case selection box, indicating that s/he wishes to select cases based upon particular values for those variables.

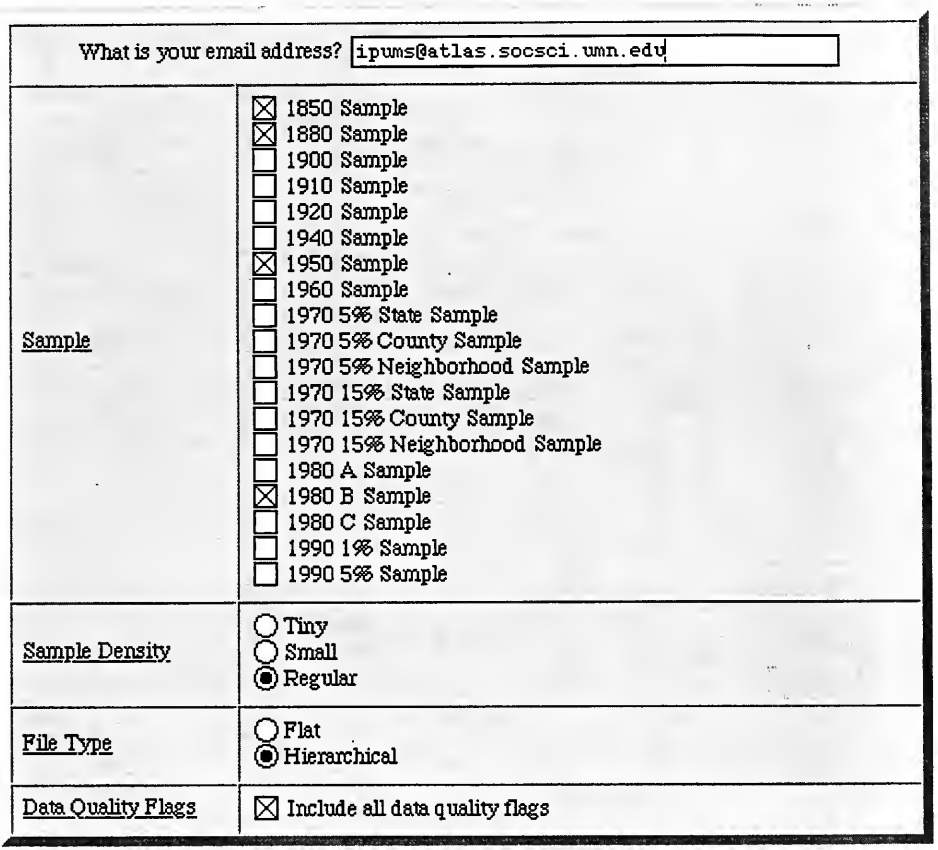


Figure 2. IPUMS Sample Selection (page 1)

☐ ELDCH	Age of eldest own child in household	x	x	x	x
☐ YNGCH	Age of youngest own child in household	x	x	x	x
☐ NSIBS	Number of own siblings in household	x	x	x	x
☐ All Select Constructed Variables					

Core Demographic Variables

Variable Name	Variable Description	Case Selection	1850	1880	1950	1980
☒ RELATE	Relationship to household head -- General		.	x	x	x
☐ *RELATE*	Relationship to household head -- Detailed		.	x	x	x
☐ IMPREL	Imputed relationship to household head		x	x	x	.
☒ AGE	Age	☒	x	x	x	x
☒ SEX	Sex	☒	x	x	x	x
☒ RACE	Race -- General	☒	x	x	x	x
☐ *RACE*	Race -- Detailed		x	x	x	x
☒ MARST	Marital status	☐	.	x	x	x
☐ AGEMARR	Age at first marriage		.	.	.	x
☐ DURMARR	Duration of current marital status		.	.	s	.
☐ MARRNO	Times married		.	.	s	x
☐ CHBORN	Children ever born		.	.	s	x
☐ All Core Demographic Variables						

Ethnicity/Nativity

Variable Name	Variable Description	Case Selection	1850	1880	1950	1980
☒ BPL	Birthplace -- General	☒	x	x	x	x
☐ *BPL*	Birthplace -- Detailed		x	x	x	x
☒ MBPL	Mother's birthplace -- General	☐	.	x	s	.

Figure 3. IPUMS Variable Selection (page 2, partial view)

The third page, shown in Figure 4, provides for case selection. Only those variables chosen for case selection on the second page will appear on the third. Depending on the type of variable, the page employs one of three procedures. For simple categorical variables such as region, the user selects values from a series of check boxes. With complex categorical variables such as birthplace, values are selected from a scroll-box that displays descriptive value labels rather than numeric codes. For numeric variables like age, users select minimum and maximum values. Users have the option of selecting: (1) only those individuals with the selected characteristics; or (2) entire households containing individuals with the selected characteristics.

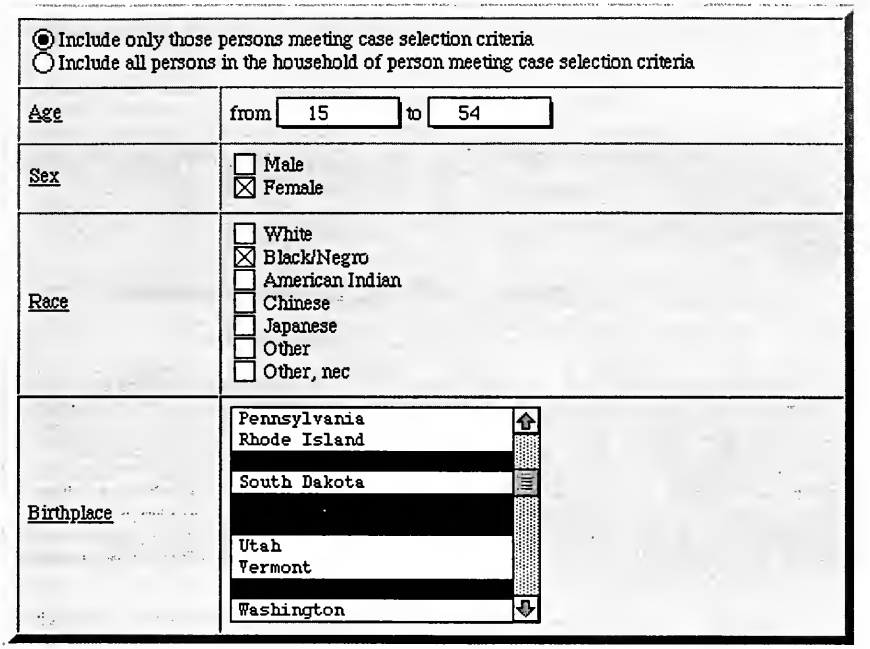


Figure 4. IPUMS Case Selection (page 3, partial view)

Extracts of the 5% samples from 1980 and 1990 are limited to cases from a single state. These files are so large that it is impractical and usually unnecessary to allow extracts on the entire samples. If either of these samples is selected, the user is forced to choose a particular state on page 3.

In Figure 4, the hypothetical user is able to select values or ranges for the variables for which the case selection box was checked in Figure 3. In this case, the researcher has chosen black women age 15 to 54 who lived in the South in the selected years. S/he has also chosen to extract only those women with the selected characteristics rather than including their entire households with them.

In the final step, users review their selections on a summary screen. If they are satisfied with their extract design, they submit it for processing. When they click the "submit" button, the program creates an extract request file that initiates the extract

engine. The engine is designed to maximize input/output efficiency. Extracts are not very demanding on the processor, but are very disk-intensive.

We will inform researchers via e-mail when their extract is completed and provide instructions for downloading their files. For each extract, users receive data, codebook, and "readme" files, and an SPSS or SAS command file. The command files will contain the column locations of variables, variable labels, value labels for categorical variables, and missing values.

b. Advantages of the Minnesota approach

A powerful set of tools associated with the World Wide Web has enabled us to reduce a complex procedure to four simple steps. Using Perl allowed us to construct dynamically each page depending on the input from the previous page. This method limits choices only to valid selections, simplifying the process for users. For example, researchers who select nineteenth century census years are not offered the options for variables on the ownership of televisions or automobiles. Our preliminary user interface can be examined at:

http://www.hist.umn.edu-/~ipums/extract.html.

There have been two previous PUMS data extraction systems developed for use on the World Wide Web. One is the Census Bureau "Data Extraction System" for the 1990 1% PUMS. The other is the University of Calgary "LANDRU" system for the 3% Public Use Microdata File of the 1991 Canadian census. These efforts represent enormous strides in data dissemination, but they are incapable of handling data of the size and complexity of the IPUMS. Because of changing variable availability across years and differing sample designs, we faced a number of design issues that the Census and Calgary systems did not.

Census Bureau Web site: **http://www.census.gov/ftp/pub/des/www/welcome.html**

Calgary Web site: **http://www.calgary.ca/~libdata/anlrud.html**

Perhaps the greatest limitation of the previous extraction systems is their inability to accommodate the hierarchical structure of PUMS data. The Public Use Microdata Samples are simultaneously samples of households and of individuals, and within households the interrelationships among individuals are known. This hierarchical structure is one of the greatest strengths of the census files. By combining the characteristics of several individuals within a household, researchers can create a wide range of new variables about family and household composition and the characteristics of family members (see "advanced extract features" below). For example, we can analyze fertility by attaching the ages of all own children to their maternal records, and we can address the family economy by simultaneously measuring the age, sex, and occupation of all family members. Neither the Census Bureau system nor that developed for the Canadian census allows users to exploit directly the information contained in the structure of the data. The Census system, for example, produces extracts of either household records or person records but not both simultaneously.

We believe our extract interface is dramatically easier to use than the Census Bureau or Canadian systems. To get a basic set of demographic variables using our extract procedure requires seven selections. By comparison, a comparable extract using the Census Bureau Data Extraction System for the 1990 PUMS requires 15 to 20 times as many selections. The Bureau system is subject to certain limitations imposed by the reliance on the SAS programming language. For example, it cannot provide information about the characteristics of persons residing with a selected subpopulation. Neither the Census Bureau nor the Calgary site accommodate more than a single census year. On the other hand, both systems are somewhat more flexible than the current IPUMS system, since they allow case selection based on the value of any variable.

c. Development of a Java-based interface

Although we believe our extract interface represents a significant improvement over existing electronic extraction software, it does have limitations. In particular, it is subject to the limitations of current Web browsers (e.g., Netscape and Mosaic), which are not truly interactive. Our present extract interface is based on dynamically produced yet static forms. Each step in the process—sample, variable, and case selection—requires a separate page, the contents of each depending on selections made on previous pages. Once a page is completed, the user cannot return to it without losing later selections. For example, if the user inadvertently omits a selection of samples or variables, s/he will have to repeat all subsequent steps. Uni-directional navigation through multiple pages inevitably complicates the procedure and increases the potential for errors.

The Internet is changing rapidly, but Java is the leading candidate to become the standard protocol for transmitting dynamic and executable content over the World Wide Web in the coming years. We plan to convert the extract interface to take

advantage of Java's unique capabilities. Java has several strengths that make it ideal for developing an intuitive and powerful user interface. The key advantage of Java for our purposes is its interactivity. There will be no need to navigate between static pages; a user's choices will mold a single page, which changes in real time. A truly interactive extract interface will reduce the risk of user error by simplifying the extract process. Essentially, when a user accesses our site, the extract interface program will be downloaded onto his/her computer. The client computer, not the server, runs the interface program which only accesses our site again to submit the extract request or to get additional information. This limits network traffic and demand on the server, and will make network connection speed less of a constraint.

The Java-based extract will also be suitable for CD-ROMs or other higher capacity random access storage media that may become available. The IPUMS is very large and pushes the bounds of what is practical with current transportable media. We plan to explore demand for a transportable version of the database (or a part of it) and pursue this avenue if warranted and feasible.

d. Advanced extract features

Over the last decade we have created thousands of specialized extracts from the PUMS using conventional higher-level programming languages. We realize that many users have special needs that go beyond the current capabilities of our preliminary extraction system. Accordingly, we plan to add several features to allow more complex subsample designs and the creation of new constructed variables. These include:

·A differential sample density feature that will allow researchers to select subpopulations at varying densities. For example, researchers might need to extract a subsample of 1-in-100 blacks and only 1-in-1000 whites in order to create the most efficient sample that would yield statistically significant results for both subgroups. The extract program will assign the appropriate weights to produce nationally representative statistics.

·A method for attaching characteristics of other household members to each individual's record. For example, labor economists often require information on the income and occupation of each individual's spouse. We plan to provide options for attaching any available characteristic of the household head, spouse of head, subfamily head, own spouse, own mother, and own father. The attached information (e.g., spouse's occupation) will appear on the person record as an additional variable.

·A method for counting the number of co-residing persons with any given set of characteristics. Some of these characteristics can define family interrelationships, permitting counts for groups within households such as unrelated persons, family members, or own children. Thus demographers using own-child fertility methods will be able to construct a set of variables giving the number of own children of each age for every mother. An economist could construct variables for the number of employed co-residing kin. The system will also be able to sum numeric characteristics (e.g., income or property) of select persons within households. This system, though complex, provides ample flexibility for advanced users.

2. Hypertext documentation

One of the greatest liabilities of the PUMS has always been the large initial time commitment associated with simply learning the organization of the documentation. The IPUMS mitigates but does not overcome this inherent weakness of conventional documentation. In order to address this problem, we intend to convert all of the IPUMS documentation into hypertext format using Adobe Acrobat "portable documentation format" (PDF). Along with the extract engine, we expect the hypertext documentation to revolutionize the way researchers use census data. The hypertext format will let users jump to relevant sections of the documentation with a simple click of a mouse.

The IPUMS simplifies the original PUMS codebooks, but there is no way to structure the documentation to eliminate the need to switch frequently between sections. Moreover, since the IPUMS consolidates eleven PUMS codebooks, its documentation is considerably more extensive than any one of them. In addition to the 800 page basic *User's Guide*, the documentation contains maps, comprehensive enumerator instructions, detailed descriptions of data transformations, and other elements that even modest users of the data would likely need to consult. By putting the data into hypertext with a generous number of links, users will be able to navigate the documentation with far greater ease than any previous PUMS codebook. They need never grapple with multiple volumes totaling 3000 pages or more.

The IPUMS will not only be the first census database to have hypertext documentation, it will be the first instance in which complete PUMS documentation is offered in any machine-readable form at all. Moreover, with hypertext, we can link the documentation directly to the extract interface so users can interactively make informed decisions when designing a sample appropriate for their research. By clicking on a variable name on the interface, the user will bring up the variable description

and comparability discussion. Tables presenting variable frequencies suggest whether particular extracts or types of analyses are feasible in a given year. Advanced users can even look up the translation tables that detail how variables were recoded from the original PUMS into their integrated format.

The hypertext documentation will be available on our Web site and can be downloaded onto a PC, Macintosh, or UNIX system. We will also make the documentation available on CD-ROM. One of the advantages of the Adobe Acrobat PDF format is its transportability across different computing platforms. We will continue to provide heavily indexed printed and word-processor versions of the documentation that will be updated to parallel any changes made in the hypertext version.

3. User support

User support is a crucial aspect of the project. The Internet and extract system dramatically increase access to the data, but one consequence is the even more urgent need to support the growing base of users. Although we designed the extract interface to be as intuitive as possible, it will still require extensive human support. Only the most recent census years are supported by the Census Bureau. There is no institutional support for any of the earlier PUMS produced by historical researchers. Only scholars at a few major demography centers are likely to get any help at all with any but the most recent samples. Undoubtedly, the lack of sustained user support has discouraged the use of the PUMS. With the advent of the Internet and e-mail, it is now possible to centralize support for all of the census samples in their IPUMS format.

Our Web site contains a hypertext e-mail link for questions concerning the extract system, data, or documentation. In the future, the Web site will refer first-time users to an on-line tutorial that will walk them through several extract examples including variable selection, case selection, and advanced constructed variable features.

The IPUMS Web site will also serve as a repository for ancillary data such as geographic contextual or occupational wage data. For example, we have already received a number of geographic boundary files from outside researchers that translate IPUMS codes into a form suitable for mapping software. In addition, members of the census project staff have developed relevant data files in the course of their own research. A central repository for such files reduces needless duplication of effort among scholars. Researchers are invited to contribute any files they may have developed that can be attached to the IPUMS. The IPUMS already contains the necessary state and county codes to link existing county-level machine-readable statistics. Thus, for example, researchers can supplement the information in the microdata with comprehensive statistics on the racial composition and average wages for the location in which an individual resided. Such multi-level analyses are increasingly a feature of historical social scientific research (Landale and Tolnay 1991; Elman forthcoming; Ruggles forthcoming). There are also city-level machine-readable data that could be adapted to correspond with IPUMS coding.

Synopsis

Electronic communication provides a unique opportunity to disseminate the census data to a much wider spectrum of the academic community than ever before. Web navigation programs like Netscape and Mosaic, e-mail, and listservers are unprecedented resources. Despite the many files downloaded from our FTP and Web sites, however, access to the IPUMS is still largely confined to major demography centers. We anticipate that with an organized effort at dissemination, combined with our project to increase accessibility, the IPUMS data will become more widely distributed than those of any individual PUMS dataset not produced as an adjunct to the most recent census. As a result, the IPUMS can be expected to resuscitate some of the historical PUMS that have hitherto been under-utilized.

Our project, as described in this paper, is a work in progress. The completion of some aspects of the extraction system, as well as our ability to make the system publicly available and to support it adequately, depends on securing additional outside funding (the status of which is pending). We expect to proceed with development even without further funding, but the pace will be slowed and the final product may involve compromises.

The IPUMS extract system and hypertext documentation will make the census samples far more user-friendly. To this point, census microdata have generally been too cumbersome for the classroom, but our project will make the data an ideal source for instructional purposes. It will be a simple matter to custom-design extracts of the appropriate size and composition for any classroom situation.

The PUMS are a unique national resource. They are the envy of researchers from other countries, but have only been partially exploited. The IPUMS project has removed many of the barriers to using the PUMS by integrating them into a single database. But the size and complexity of the database still pose formidable obstacles to access and usability. Our project addresses both these problems, enabling researchers with very limited computing resources to take advantage of the IPUMS.

References

Elman, Cheryl (forthcoming) "Old Age, Economic Activity, and Living Arrangements in the Early 20th Century United States." *Social Science History.*

Farley, Reynolds, and William Frey (1994) "Changes in the Segregation of Whites from Blacks." *American Sociological Review* 59: 23-45.

Gjerde, Jon, and Anne McCants (1995) "Fertility, Marriage, and Culture: Demographic Processes Among Norwegian Immigrants to the Rural Midwest." *Journal of Economic History* 55: 860-888.

Gordon, Linda, and Sara McLanahan (1991) "Single Parenthood in 1900." *Journal of Family History* 16: 97-116.

Graham, Stephen N. (1980) *1900 Public Use Microdata Sample User's Handbook.* Seattle: Center for Demography and Ecology, University of Washington.

Haines, Michael (1989) "American Fertility in Transition: New Estimates of Birth Rates in the United States, 1900-1910." *Demography* 26: 137-148.

Hirschman, Charles, and Ellen P. Kraly (1990) "Racial and Ethnic Inequality in the United States, 1940 and 1950: The Impact of Geographic Location and Human Capital." *International Migration Review* 24: 4-33.

Jacobs, Jerry A. (1989) "Long-Term Trends in Occupational Segregation by Sex." *American Journal of Sociology* 95: 160-173.

Jenson, Leif (1991) "Secondary Earner Strategies and Family Poverty: Immigrant-Native Differentials, 1960-1980." *International Migration Review* 25: 113-140.

Johnson, N., and S. Lean (1985) "Relative Income, Race and Fertility." *Population Studies* 39: 99-112.

Kalmijn, Matthijs (1994) "Assortative Mating by Cultural and Economic Occupational Status." *American Journal of Sociology* 100: 422-452.

Krivo, Lauren (1995) "Immigrant Characteristics and Hispanic-Anglo Housing Inequality." *Demography* 32: 599-615.

Landale, Nancy S., and Stewart Tolnay (1991) "Group Differences in Economic Opportunity and the Timing of Marriage: Blacks and Whites in the Rural South 1910." *American Sociological Review* 56: 33-45.

Magnuson, Diana (1995) "The Making of a Modern Census: The United States Population Census, 1790-1950." Ph.D. Dissertation, University of Minnesota.

Mare, Robert D. (1991) "Five Decades of Educational Assortative Mating." *American Sociological Review* 56: 15-32.

Morgan, S. Philip, Antonio McDaniel, Andrew T. Miller, and Samuel Preston (1993) "Racial Differences in Household Structure at the Turn of the Century." *American Journal of Sociology* 98: 798-828.

Olson, Thomas (1991) "The Women of St. Luke's and the Evolution of Nursing, 1892-1937." Ph.D. Dissertation, University of Minnesota.

Ruggles, Steven (1994a) "The Origins of African-American Family Structure." *American Sociological Review* 59: 136-151.

Ruggles, Steven (1994b) "The Transformation of American Family Structure." *American Historical Review* 99: 103-128.

Ruggles, Steven (forthcoming) *Fragmentation of American Family Structure, 1850-1990.* Cambridge, MA: Harvard University Press.

Ruggles, Steven and Russell R. Menard (1994) "Public Use Microdata Sample of the 1880 United States Census of Population: User's Guide and Technical Documentation. (Inter-University Consortium for Political and Social Research).

Ruggles, Steven, Russell R. Menard, Lisa Dillon, and Matt Mulcahy (1995) "1850 Public Use Microdata Sample: User's Guide." (Inter-University Consortium for Political and Social Research).

Ruggles, Steven and Matthew Sobek (1995) *Integrated Public Use Microdata Series: User's Guide* (University of Minnesota, Social History Research Laboratory).

Sanderson, Warren (1987) "Below-Replacement Fertility in Nineteenth-Century America." *Population and Development Review* 13: 305-313.

Sandefur, Gary D., and Arthur Sakamoto (1988) "American Indian Household Structure and Income." *Demography* 25: 71-80.

Sassler, Sharon (1995) "Trade-Offs in the Family: Sibling Effect on Daughters' Activities in 1910." *Demography* 32: 557-575.

Shoemaker, Nancy (1991) "The American Indian Recovery: Demography and the Family, 1900-1980." Ph.D. Dissertation, University of Minnesota.

Sorenson, Ann Marie (1989) "Husbands' and Wives' Characteristics and Fertility Decisions: A Diagonal Mobility Model." *Demography* 26: 125-135.

Strong, M. A., et al. (1989) *User's Guide Public Use Sample 1910 United States Census of Population.* Philadelphia: Population Studies Center, University of Pennsylvania.

U.S. Bureau of the Census (1972) *Public Use Microdata Samples of Basic Records from the 1970 Census: Description and Technical Documentation.* Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1973) *Technical Documentation for the 1960 Public Use Microdata Sample.* Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1982) *Public Use Microdata Samples of Basic Records from the 1980 Census: Description and Technical Documentation.* Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1984a) *Census of Population, 1940: Public Use Microdata Sample Technical Documentation.* Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1984b) *Census of Population, 1950: Public Use Microdata Sample Technical Documentation.* Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1992) *Census of Population and Housing, 1990: Public Use Microdata Sample U.S. Documentation.* Washington, D.C.: U.S. Government Printing Office.

Watkins, Susan, ed. (1994) *After Ellis Island: Newcomers and Natives in the 1910 Census.* New York: Russell Sage.

1 Submitted for the IASSIST Conference held in Quebec, Canada. May 1995

2. Send correspondence to: IPUMS University of Minnesota Department of History, 614 SST 267 19th Ave S. Minneapolis, MN 55455 ipums@hist.umn.edu (612) 624-5818

Democratic Elections on the Internet: The Lijphart Elections Archive

by Renata G. Coates[1] ,
University of California, San Diego

The Archive
The Lijphart Elections Archive, housed at the University of California, San Diego campus, is a research collection of district-level election results for twenty-seven countries. Until 1994, the collection focused on post-World War II democracies in Western Europe, but included the United States, Canada, India, Israel, Japan, Australia and New Zealand. Recently, Costa Rica and the European Union were added. Future plans call for expansion of the Archive to more than 70 countries—including many new democracies from Central and Eastern Europe, Latin America, and Africa.

The collection includes all national elections for the lower house (in some cases the only house) of the legislature. Where the legislature is bicameral, the upper house is included only if it is directly elected by the voters. The volumes in the collection are the detailed, district-level election results that are usually published by government statistical offices in one or more volumes for each election. In some cases, non-governmental publications are included if they are the more complete source.

The data includes the number of votes by party for each election district, including detailed lists for minor and major parties, and the number of seats or "elected candidates" by party for each election district. Complete election data was sought for those countries where there is preferential voting (e.g., Australia and Ireland), where there is a first or second ballot (e.g., France), and where there is proportional representation (e.g., where one candidate may be elected from several districts).

Until 1994, the preferred format for building this Library collection had been the original hard copy. During the last year, prompted by requests from graduate students, efforts have been made to acquire data in machine-readable form.

Historic Origins
The Lijphart Elections Archive is named for Arend Lijphart, Research Professor of Political Science at the University of California, San Diego, and the man responsible for its establishment. 2 Prof. Lijphart, is a world-renowned expert on elections.

When he began researching comparative electoral systems back in 1984, no library anywhere collected all the detailed statistical data that he needed. After consulting with international colleagues as to the need for such an archive, Professor Lijphart and the University Library agreed to create such a collection at UCSD. The task of finding materials that in many cases was already out of print, proved to be difficult at times. Much of the success of the collection must be attributed to Professor Lijphart himself. Many of the volumes in the collection were acquired and donated by him. It should also be noted here, that the collection was named in his honor by the University's librarians.

Present and Future
The establishment of the Internet and the emergence of many more democracies has infused a new energy into the Lijphart Archive. Although adding machine-readable data to the collection was always a consideration, it wasn't until a critical mass of technology, people and funding formed, that it became a reality. New people have now joined the team. Jim Jacobs as the University Library's Data Librarian and I, as the Political Sciences Bibliographer along with Professors Gary Cox and Matthew Shugart, of the Political Science Department, hope to enlarge both the content and the access of what we see to be a worldwide resource for scholars. Others apparently share our vision, since establishing the Lijphart Elections Archive as a permanent Internet web site has recently received funding from the National Science Foundation. One result of this initial funding has been the creation of a home page in March of 1995. 3

Additional funding will be sought from NSF and other sources. Enlarging the Archive to include pre-1994 data has received some discussion. The possibility of scanning existing paper holdings is being considered. OCR reliability, the condition of the original materials, the different languages and type fonts and lastly, cost, are some of the concerns that will need to be dealt with.

Web Site

The current home page for the Lijphart Elections Archive is still under construction and is part of UCSD's Social Sciences Data Collection. Presently, the Archive's holdings can be searched in a number of different ways: by material type (eg. paper holdings, vs. electronic holdings), by country, by electoral system and by keywords. The program can also show the library catalog entry for the item— a feature that provides information suitable for generating an interlibrary loan transaction. In addition to providing the actual data, the Lijphart home page will link users to other election information on the Internet. A current example, is Slovakia. The Lijphart Archive never collected election data on Slovakia, but since the information has been made available through EUNET, we have added a linkage for the convenience of our researchers.

Conclusion

The UCSD Social Sciences and Humanities Library is still actively acquiring election returns in hard copy. Although our plans are to expand the Archive's holdings to all the worlds' democracies, much of our success will depend on funding from outside the University. Until we are assured of the stability of the Internet site, we will probably be duplicating the data in hard copy—to maintain the integrity of the Lijphart Archive for future researchers.

Plans are underway to submit another proposal to the National Science Foundation. Simultaneously, efforts are being pursued for international agreements with other agencies for sharing both information and more important, staff and data. Creators of the home page envision it as a major resource for locating and reflecting other Internet election resources. The on line version of the Lijphart Election Archive however, will continue to have the same goal as the paper collection —to provide researchers with the actual data of elections down to the district level. We know this is an ambitious project, but we also have come to realize that promoting the results of democratic elections worldwide is a worthwhile effort for all of us.

1 Submitted for the IASSIST Conference held in Quebec, Canada. May 1995

2 Prof. Lijphart served as President of the American Political Science Association during 1994-95 and is the author of Electoral Systems and Party Systems: A Comparative Study of Twenty-Seven Democracies, 1945-1990 Oxford: Oxford University Press, 1994.

3 Lijphart Elections Archive URL: <http://ssdc.ucsd.edu/lij>

Establishing Data and Documentation Standards for Investigators who are Required to Archive Research Data

by Patrick T. Collins[1], Project Director, National Data Archive on Child Abuse and Neglect

Introduction

This paper is about the approach that the National Data Archive on Child Abuse and Neglect has taken to improving the quality and consistency of our documentation. Of the many problems we have encounytered, including uncooperative investigators, dirty data files, and unusual file formats, poorly prepared or non-existent documentation has been the most difficult to handle. Since investigators were not required to archive their data with our Archive, we had to actively solicit contributors. Most researchers were unwilling to contribute their data and the ones who were willing had little or no resources to dedicate to the task. In short, we were in the position of having to accept whatever investigators were willing to provide. In many cases we received nothing more than an SPSS or raw data file and a copy of the instrument, leaving us with the daunting task of creating a user's guide from scratch. The task of preparing comprehensive documentation for these studies was so time consuming that were only able to process 2-3 datasets per year. While we appreciated the efforts of the investigators who chose to contribute data to the Archjive, it became clear that the only way the Archive could expand its holdings with any speed would be to improve the nature of the materials contributed by investigators. Since our Archive was funded by a federal agency with an active research program, we chose to work through that agency in order to establish data documentation standards for their research grantees. But before I describe this process in detail, let me tell you a little more about the Archive.

The National Data Archive on Child Abuse and Neglect

The National Data Archive on Child Abuse and Neglect has been in operation for approximately six years. During this time NDACAN has received all of its funding from the National center on child Abuse and Neglect (NCCAN) which is a division of the Administration for Children Youth and Families which in turn is a unit of the US Department of Health and Human Sercies. NCCAN is the federal agency witht the primary mission of responding to child abuse and neglect in the USA. One of NCCAN's many responsibilities is a field initiated research program which is funded at approximately $1.5 million per year. The Archive, which is funded through this program, works primarily with NCCAN's research grantees. We are in the final year of our second three-year award from NCCAN and will apply this spring for continued funding. Over its six years of operation, the Archive has been flat-funded at $150,000 per year, leaving us with approximately $100,000/year in direct funds. Most of these funds are used to support our 2.2 FTEs.

The Archive's primary mission is to acquire, process, preserve, and disseminate high quality datasets relevant to the study of child maltreatment. We have the secondary mission of networking and training child maltreatment workers. Toward this end, the Archive publishes a biannual newsletter, hosts a listserv with approxixmately 400 subscribers, and maintains a Gopher/FTP server. In many ways we have been more successful in achieving our secondary mission of networking and training researchers than our primary mission of acuirring datasets. creating networking and training opportunities is a fairly straightforward job and such services are eagerly consumed by researchers. Acquiring, processing, and disseminating high quality data is a far more complex task.

For these and other reasons, NDACAN began to advocate for mandatory data archiving for NCCAN research grantees. Simultaneously, we lobbied NCCAN to establish tecnical standards for their research grantees. Toward this end, Jane Powers and I co-authored, The Preparation of Data Sets for Analysis and Dissemination: Technical Guidelines for Machine-Readable Data, a manual which set forth standards for the preparation of research datasets and their associated documentation. We disseminated hundreds of copies of this manual to CNNAN's research grantees and to other child maltreatment researchers and we offered technical assistance to researchers willing to follow our guidelines. Unfortrunately very few researchers responded with interests. The situation changed however when, in their 1993 RFP, NCCAN announced that their research grantees would be expected to prepare datasets and documentation according to NCACAN's standrads. This generated quite a bit of interest and for the first time applicants began to contact us for technical assistance and copies of our manual.

The Archive's lobbying efforts came to fruition when, in their 1994 RFP, NCCAN set forth the requirement that applicants include in their proposal plans to prepare their data and documentation according to NDACAN's guidelines and to archive

their data with NDACAN upon the completion of their grant. As a result of this dramatic policy change, NDACAN will have the opportunity work with investigators from the beginning of their projects to ensure that data and documentation are prepared properly. all grantees will be provided with free technical assistance during the start-up phase of their studies and will receive a new publication entititled, Depositing Data with the National Data Archive on Child Abuse and Neglect: A Handbook for Investigators. The purpose of the handbook is to outline the investigators responsibilities and to provide a clear set of deliverables that must be submitted to NDACAN.

In some sense NCCAN's policy change took us by surprise. After years of lobbying, we were happily surprised to learn that NCCANdecided to require research grantees to archive their data. The way the requirement was implemented was that NCCAN reviews rated applicants' plans to prepare and archive data and documentation as one of the many criteria use to evaluate grant proposals. While it is not clear that this arrangement provides any method of enforcement, grantees are working under the assumption that they will be required to archive their data. While NCCAN sèt forth the requirement, the Archive is in the position of defining all of the nuts and bolts of the arrangement. Instead of reinvention the wheel, we have studied the data archiving programs of other federal agencies, such as the national Institute of Justice (NIJ) and the National Science Foundation (NSF). Our approach has been to build on the successes of these programs and make adjustments where necessary. Our goal is to build a program that meets the needs of NDACAN and NCCAN's research grantees.

In our experience of working with researchers, their greatest concern is having adequate time to publish their results of their study before the data are released to the public. In response, we have created a policy that will allow all investigators a two-year "grace period" after the termination of their grant which will allow them to publish the results of their study before the data are made available to the public. This is an area where our approach differs from that of NIJ. NIJ grantees must submit their data and documentation along with their final report at the termination of the award; grantees who fail to do so are not eligible for new NIJ funding. This policy has created a great deal of animosity among some NIJ grantees and there have been cases where a secondary data user published a study's findings before the principal investigator.

In other areas, we have closely followed NIJ's lead. For example, in determining the investigators' responsibilities and required deliverables we have essentially mimicked NIJ's requirements. Broadly defined, we see the investigators are responsible for, submitting data and sporting materials, responding to requests by NDACAN staff for additional or clarifying information, and reviewing and correcting draft materials prepared by the NDACAN staff.

NDACAN staff is responsible for preparing the data files in ready-to-use statistical file formats, preparing a user's guide that describes the project and data, reviewing the codebook for completeness and accuracy and augmenting the codebook as necessary, and making copies of the datasets archived available to the research community (for a small fee) and providing technical support to data users.

This new arrangement has the potential to solve the problem of inadequate and inconsistent documentation because we can specify exactly what materials the investigator must submit. Grantees will be provided with a clear set of deliverables as well as clear written guidelines for the preparation of those materials. Working with grantees early on in their projects in order to determine potential problem areas and needs for technical assistance will be integral to our approach.

While it will be several years before the first grantees are required to submit data under this arrangement we have established a tentative list of deliverables. These include:

(1) Data file(s).
(2) Description of data files
(3) Data collection instruments(s)
(4) References for data collection instruments
(5) Codebook or data dictionary
(6) Explanation of derived (computed) variables
(7) Final project report, project summary, or other description of the project
(8) Bibliography of publications pertaining to the data
(9) Printout of the first and last data records

The draft handbook that I have distributed contains some guidelines and specifications for the preparation of these materials. Our plan is to distill the most important guidelines in our technical standards manual and include them in the handbook. We want to keep the handbook as simple and free from jargon as possible. Once it is completed, we will stop distributing the technical standards manual because it is both out-dated and too technical for investigators. We are very interested in your

feedback and suggestions for the handbook so please read it if you have time. And let us know how we can improve it.

Once we receive these materials from the investigators we will create a comprehensive user's guide with the following format:

Project Overview
- Purpose of the Study
- Sampling/Selection Information
- Data Collection
- Instruments and Measures

Description of Machine-Readable Files
- List of Files
- Notes Regarding the Data Files

References
- References to publications from the dataset
- References to publications related to the dataset

Appendices
- Data Collection Instruments
- Codebook
- Sample Programs

So far reaction to the policy has been relatively positive. We presented our general plan to a large group of researchers at the annual NCCAN grantees meeting and most of the grantees were receptive. We are in the process of forming an advisory committee to handle cases where investigators have special needs relative to archiving (e.g., longitudinal studies). We are hopeful that NCCAN's data archiving policy will go a long way toward solving the problems I have described however, it will be several years before we know for sure. Clearly, the approach has limitations but we feel it is a step in the right direction.

1. Submitted for the IASSIST Conference held in Quebec, Canada. May 1995

CALL FOR PAPERS

Global Access, Local Support: Social Science Computing in the Age of the World Wide Web

The **International Association for Social Science Information Service and Technology** (IASSIST) and the **Social Science Computing Association** (SSCA) announce their joint 1998 ***Conference, Global Access, Local Support: Social Science Computing in the Age of the World Wide Web.*** The conference will be held May 19-22 on the Yale campus in New Haven CT and will address social science computing and information services in the research, teaching, and data library arenas. This is IASSIST's 24th annual conference, and the 9th SSCA Computing for the Social Sciences conference.

The joint program committee welcomes proposals for papers, panels, poster sessions and discussions that focus on the conference theme of providing local support for global access. Themes that may be woven into the conference include: new avenues for academic teaching; quantitative data in the age of the Internet; non-traditional data analysis, presentation, and support; training and ongoing support for the end user; technical aspects of data base management and Web delivery; metadata standards, resource discovery; social implications of global networking; economic and legal implications in international perspective. In addition to sessions devoted to these specific themes, there will be opportunities to present other topics of interest to the membership of the two sponsoring organizations. Ample opportunity for small group interaction and the exchange of ideas will be encouraged.

PROPOSALS ARE DUE JANUARY 5, 1998. Notification of proposal acceptance will be made by February 13, 1998, sent by E-mail or post.

AUDIENCE: IASSIST conferences bring together professionals who are engaged in the creation, acquisition, processing, documentation, maintenance, distribution, preservation and use of computerized social science data. The Social Science Computing Association brings together users and designers of computer-based systems for social science research, including educators and students interested in the links between the social sciences and technology. The complimentary and overlapping interests of these professional organizations provide a rich spectrum of expertise in which to explore new ideas and solutions.

CONFERENCE THEMES: Access for the individual user to both numeric and textual data has increased exponentially through services offered via the Internet, in particular through the World Wide Web. The role of libraries and data archives has expanded from local holder of physical collections to gateway to a multitude of information providers and to being global providers of information. For faculty and students, the Internet offers opportunities for instructional innovation, additional channels of communication, and access to data traditionally believed to be inaccessible. This information explosion raises a number of legal, economic, archival, administrative, and technical questions for users, producers, and service providers. In addition, ease of access for the end user produces both opportunities and challenges for local instructional and research support. Join IASSIST and SSCA as we explore challenges and strategies to put new technology to optimal use, to create a structured and functional 'global village' rather than a chaos of information overflow.

Within the general theme, we will focus on several subthemes including:

New avenues for academic teaching

- Bringing primary research into the classroom
- Instructional technology (CD-ROM, multimedia, Internet)
- Distance Learning
- Teaching qualitative methods in a network environment
- Push technology in the classroom
- Computer literacy standards for students and faculty
- Electronic reserves, course materials on the WWW

Quantitative data in the age of the Internet

- Data libraries in the age of FTP
- Wholesale delivery to statistics on demand
- Interfacing web form based user demands and statistical software
- Limits of "remote data analysis"
- Improving access to policy and public opinion data

Non-traditional data analysis, presentation, and support
- Revitalization of textual (document) analysis
- Graphical presentation; scientific visualization; spatial analysis
- Software for computer-assisted content analysis

Training and ongoing support for the end user
- Who has the responsibility? what are reasonable expectations?
- Training models and materials; statistical laboratories
- Changing job profiles to cope with changing technologies

Technical aspects of data base management and Web delivery
- Integrating data and documentation in WWW based data systems
- Interfacing multi-purpose data bases and Web servers
- Facilitating user access (response time, transmission speed, etc.)
- Converting print media to electronic storage, migrating and preserving digital information

Metadata standards, resource discovery
- Metadata format options (HTML, PDF, RTF, SGML, etc.)
- Metadata content standards, the Data Documentation Initiative
- Bringing social science documentation into the digital age
- Locating resources: thesauri, indexing, Internet resource discovery systems
- Data documentation on the WWW

Social Implications of Global Networking
- MUDs, MOOs, and personal identity
- Long distance scientific collaboration over the Internet
- Ethical issues for scientific research on computer networks

Economic and Legal Implications in International Perspective
- "Internet for free" vs. "Internet for fee"
- Authors' royalties and publishers' profits vs. limited library budgets
- Copyright and fair use, privacy, confidentiality, security
- Government data and access fees
- Assessing quality and validity of information via the Web

CONFERENCE DETAILS: The conference will open with workshops on Tuesday, May 19th, followed by 2 _ days of plenaries, sessions, panels, lunch speakers, and social events. The Conference sessions will be held on the beautiful and historic campus of Yale University. Hotel accommodations are within comfortable walking distance of the conference site. A final weekend in New York City can be arranged as a supplemental option.

FEES AND REGISTRATION: Full registration materials will be mailed to IASSIST and SSCA members, will be posted on numerous listservs, and will be available on the www site. All those submitting proposals will receive complete registration and accomodation information. All paper presenters must register and pay the registration fee.

CONFERENCE PROCEEDINGS: The papers presented at the conference will be published by IASSIST and SSCA. Presenters will be asked to provide a formal paper for publication in either the IASSIST Quarterly or a special edition of Social Science Computing Review. Print and electronic versions of papers are encouraged.

SUBMISSION GUIDELINES: We seek proposals for papers, poster sessions, panel discussions, demonstrations, and workshop presentations. Proposals may be submitted electronically, forms are available at the Conference WWW site:

`http://www.columbia.edu/acis/eds/iassist98/`

ADDITIONAL INFORMATION:

Information about IASSIST may be found:
http://datalib.library.ualberta.ca/iassist

Information about the Social Science Computing Association may be found:
http://ag.arizona.edu/ssca/

Contact the Yale organizers for conference information:

Ann Green, Yale Social Science Statistical Laboratory
PO Box 208208 New Haven,
CT 06520-8208
(203) 432-3277 FAX (203) 432-6976
email: ann.green@yale.edu

Jocelyn Tipton, Yale Social Science Library
PO Box 208263
New Haven, CT 06520-8263
(203) 432-3310 FAX (203) 432-8979
email:jocelyn.tipton@yale.edu

CONFERENCE INTENTION FORM

Fax or mail this form before January 5, 1998
Forms can also be found at:
http://www.columbia.edu/acis/eds/iassist98/

SEND TO:
Yale Statlab
PO Box 208208
New Haven, CT 06520-8208
fax (203)432-8979
phone (203)432-3277
email ann.green@yale.edu

Name _____
Title _____
Affiliation _____
Mailing Address _____
Email _____
Phone FAX _____

I intend to submit a paper with the following title: _____

_____

I am interested in presenting the following poster session/demonstration: _____

_____

I will organize a session on: _____

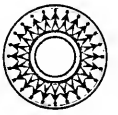
_____

Please include a 250-500 word abstract plus a 2 sentence biographical summary with the above.

I am willing to chair a session. _____

National Data Archive on Child Abuse and Neglect
Summer Research Institute

Cornell University, June 14-19, 1998



The National Data Archive on Child Abuse and Neglect will sponsor its annual Summer Research Institute (SRI) on June 14-19, 1998. The Institute provides a unique opportunity for scholars to conduct secondary analyses in the field of child abuse and neglect. Participants represent a wide variety of disciplines and are selected on a competitive basis. The primary goals of the Institute are to provide training to child abuse and neglect researchers and to increase the amount of scholarly work conducted with the Archive's holdings. In addition, the Institute provides child abuse and neglect researchers an invaluable opportunity for networking and collaborating with each other.

Scholars, professionals involved in research, and advanced graduate students are all encouraged to apply. Fifteen applicants will be selected based on their previous research experience and level of commitment to following their work through to publication.

Applications are available from
The National Data Archive on Child Abuse and Neglect
Phone: (607) 255-7799,
WWW: http://www.ndacan.cornell.edu

Applications must be received by February 15, 1998

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICE AND TECHNOLOGY

• • • •

ASSOCIATION INTERNATIONALE POUR LES SERVICES ET TECHNIQUES D'INFORMATION EN SCIENCES SOCIALES

The **International Association for Social Science Information Services and Technology** (IASSIST) is an international association of individuals who are engaged in the acquistion, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompases hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the IASSIST QUARTERLY. They also benefit from reduced fees for attendance at regional and international conferences sponsored by IASSIST.

Membership fees are:
Regular Membership. $40.00 per calendar year.
Student Membership: $20.00 per calendar year.

Institutional subcriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subcription:
$70.00 per calendar year (includes one volume of the Quarterly)

Membership form

I would like to become a member of IASSIST. Please see my choice below:

- ☐ $40 Regular Membership
- ☐ $20 Student Membership
- ☐ $70 Institutional Membership

My primary Interests are:

- ☐ Archive Services/Administration
- ☐ Data Processing
- ☐ Data Management
- ☐ Research Applications
- ☐ Other (specify)_____

Please make checks payable to IASSIST and Mail to :
Mr. Marty Pawlocki
Treasurer, IASSIST
% 303 GSLIS Building,
Social Science Data Archives, University of California, 405 Hilgard Avenue, Los Angeles, CA 90024-1484

Name / title

Institutional Affiliation

Mailing Address

City

Country / zip/ postal code / phone